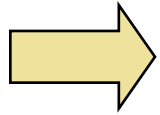




Ultrapedia: An Analytical Encyclopedia

**Jesse Wang
Project Halo, Vulcan Inc.**



■ **Background: Project Halo and Wiki Technology**

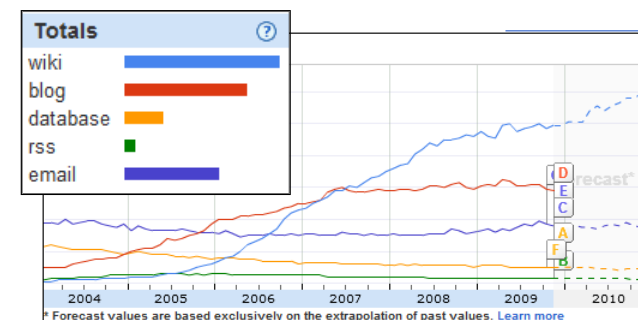
- Problem: Building Large Science Knowledge Bases
- Crowdsourcing Data: Semantic MediaWiki
- An Early Attempt: Semantic Sci-Fi Movie Wiki

■ **Ultrapedia**

- A Vision of a Semantic Encyclopedia
- Ultrapedia Under the Hood
- Ultrapedia Demo
- The Future

Project Halo and Wikis

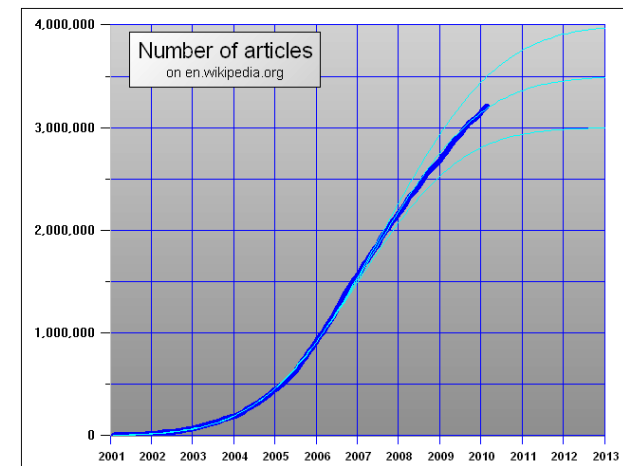
- **Project Halo is a staged R&D effort to build very large, computer-based question answering systems**
 - Initially in natural science, to answer questions at the AP level
 - Traditional search systems (Google/Yahoo/Microsoft, Endeca) do not work for this task
- **Three Project Halo challenges:**
 - Knowledge Formulation: Can you build the knowledge bases?
 - Question Formulation: Can you query the knowledge bases?
 - Question Answering and Explanation Generation: Can you get the answers?
- **Knowledge Formulation Challenge: create technology to build very large, computer-processable knowledge bases**
 - Millions of interlinked assertions, rules, and patterns, structured to support for question-answering algorithms, built and maintained in a cost-effective, reliable way
- **Wiki is a popular way to crowd-source documents**
 - Highly reliable, Internet-scale, and incredibly cheap



Can we use wiki techniques to build large knowledge bases?

A useful wiki - Wikipedia

- Wikis are tools for textual *Publication* and *Consensus*
- MediaWiki is the software for Wikipedia, Wikimedia, Wikibooks, etc.
 - Most successful Wiki package
 - High performance: 10K pages/sec served, scalability demonstrated
 - LAMP web server architecture, GPL license
 - 3 major data centers, 6-8 fulltime technical employees
 - Publication: simple distributed text authoring model
 - Wikipedia: >3.2M English articles, >270M edits, >2.5M images, #8 Alexa traffic rank in January
 - ~8.5M user accounts; ~157K active in the last month
 - 264 languages (English and German are top two)
 - Consensus achieved by global editing and rollback
 - ~17 edits/page on average (with high variance)
 - Fixpoint hypothesis, although consensus is not static
 - Gardener/admin role for contentious cases



English Wikipedia Article Counts

- Wikipedia error rates are comparable to Encyclopedia Britannica
 - http://en.wikipedia.org/wiki/Wikipedia:External_peer_review



Wikipedia is Fantastic, but...

- **Most Wikipedia data cannot be easily exploited by computers**
 - Searching is limited to keywords only, and the result unit is the whole article
 - No easy way to extract from Wikipedia, for example:
 - Sci-Fi movies after Y2K that cost <\$10M while gross >\$30M
 - All Porsche models that accelerate less than 6 seconds
 - Skyscrapers in China higher than 50 stories, built before 2000
 - Soccer player with jersey #11 from a club with a home stadium with more than 40000 seats, who were born in a country with more than 10M people
- **Tables in articles are (mostly) manually built**
 - By definition out of date, often inconsistent with text
- **Category hierarchy is inconsistent and arbitrary**
 - “Page:List_of_Rivers_of_Alabama” is not related to “Category:Lists_of_Rivers_of_the_United_States”

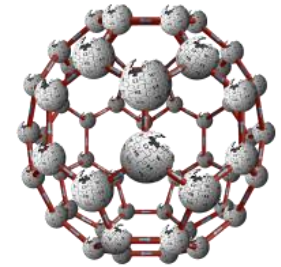
Enter Semantic Wikis!

Blend the *ease* and *consensus* of a wiki
with the *query power* of a database



How can Wikipedia be enhanced?

Sci-Fi Movie Wiki Demo



■ A Semantic Wiki for Science Fiction Movies from 2000 onward

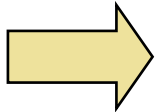
- Look and feel are “classic” MediaWiki
- Pages came from Wikipedia
- Extracted semantic data from the infoboxes (via patterns, automatic)
- Semantic UI enhancements
 - Factbox
 - Pivoting / Browsing
 - Web Services Integration
 - Sidebar Tree View

■ Things to take away:

- Data is now part of the wiki consensus process
- Data can be exploited in articles
 - For mashups to other web data (Amazon in this example)
- Data enables enhanced usability relative to Wikipedia

■ Background: Project Halo and Wiki Technology

- Problem: Building Large Science Knowledge Bases
- Crowdsourcing Data: Semantic MediaWiki
- An Early Attempt: Semantic Sci-Fi Movie Wiki



■ Ultrapedia

- A Vision of a Semantic Encyclopedia
- Ultrapedia Under the Hood
- Ultrapedia Demo
- The Future

What Would a Powerful Encyclopedia Look Like?

- **Idea: Wikipedia articles merged with the power of a database – enabling analysis of data**
- **For Authors: tools to create more compelling articles**
 - Great visualizations: charts, tables, timelines, photos, analytics
 - Always up-to-date across the Encyclopedia
 - Encourage data consistency and find data errors
 - Link in other web data sources
- **For Readers:**
 - Enhanced articles and data interaction
 - Faceted navigation
 - Sophisticated queries (both standing and ad-hoc)
 - Integrating and mining data inside and outside of Ultrapedia



Problem: How do you keep the data up to date?

- Leverage the live stream of updates from millions of Wikipedia authors
- Data is embedded in the article text, with simple ways for article authors to maintain and extend it. No DBAs.
- Authors and readers always in the loop for merging, updating, validating, mapping
- Data is reviewed just like text is reviewed

Can't do it on Wikipedia directly...



Ultrapedia : Wikipedia + a little semantics



■ Goal: Prototype a small semantic encyclopedia

- Create an semantic version of a part of Wikipedia
- Software is SMW and some Halo extensions
- Wikipedia-based checking and corrections

■ Ultrapedia Prototype Details

- Test domain is German cars
- ~2500 Wikipedia pages, ~40000 triples
- Private versions of Wikipedia, SMW, OB, and DBpedia hosted at wiking.vulcan.com
- Features
 - Corrections flow from Wikipedia to Ultrapedia in real time
 - Full data source tracking from Wikipedia
 - Wikipedia table ingestion and parsing
 - Feedback (user rating) loop for data
 - New visualizations for tables, charts, photos
 - External data integrated into articles
 - SPARQL-based queries
 - Derived assertions (via OntoBroker)

| Class | Articles | Infobox Articles |
|-----------------|-------------|------------------|
| Company | 134 | 53 |
| Person | 93 | 57 |
| Automobile | 370 | 345 |
| Auto Generation | 1480 | 1380 |
| Engine | 135 | 12 |
| Other | 283 | 3 |
| Totals | 2495 | 1850 |



Extracting Structured Data from Wikipedia



WIKIPEDIA
The Free Encyclopedia

Navigation

- Main Page
- Contents
- Featured content
- Current events
- Random article

Interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

Search

Go Search

Box

What links here

Related changes

Upload file

Special pages

Printable version

Permanent link

Cite this page

Languages

- Anglo Saxon
- العربية
- Български
- Català
- Česky
- Cymraeg
- Dansk
- Deutsch
- Eesti
- Español
- Esperanto
- Euskara
- Français
- हिन्दी
- Ido
- Bahasa Indonesia
- Íslenska
- Italiano
- עברית
- Latviešu
- Lëtzebuergesch
- Lietuvių
- Magyar
- Nederlands
- 日本語
- Norsk (bokmål)
- Norsk (nynorsk)

Bristol

From Wikipedia, the free encyclopedia

Coordinates: 51°28′N 2°35′W﻿ / ﻿51.467°N 2.583°W﻿ / 51.467; -2.583

This article is about the English city. For other uses, see [Bristol \(disambiguation\)](#).

Bristol (ⓘ pronunciation (help·info); IPA: ˈbrɪstəl) is a city, unitary authority and ceremonial county in South West England, 105 miles (169 km) west of London, and 44 miles (71 km) east of Cardiff.

With an approximate population of 410,950, and urban area of 550,200, it is England's sixth, and the United Kingdom's ninth most populous city, one of England's *core cities* and the most populous city in South West England. It received a royal charter in 1155 and was granted county status in 1373. For half a millennium it was the second or third largest English city, until the rapid rise of Liverpool, Birmingham and Manchester in the Industrial Revolution in the later part of the 18th century. It borders on the Counties of Somerset, and Gloucestershire, between the cities of Bath, Gloucester and Newport, and has a short coastline on the estuary of the River Severn, which flows into the Bristol Channel.

Bristol is one of the centres of culture, employment and education in the region. From its earliest days, its prosperity has been linked to that of the Port of Bristol, the commercial port, which was in the city centre but has now moved to the Severn estuary coast at Avonmouth and Portbury. In more recent years the economy has been built on the aerospace industry, and the city centre docks have been regenerated as a centre of heritage and culture.^[2]

Contents [hide]

- 1 Boundaries
- 2 History
- 3 Economy and industry
- 4 Culture
 - 4.1 Arts
 - 4.2 Sport and leisure
 - 4.3 Media
 - 4.4 Dialect
- 5 Politics and government
- 6 Demographics
- 7 Physical geography
- 8 Education, science and technology
- 9 Transport
- 10 Twin cities
- 11 See also
- 12 References
- 13 External links

Boundaries

There are a number of different ways in which Bristol's boundaries are defined, depending on whether the boundaries attempt to define the city, the built-up area, or the wider "Greater Bristol". The narrowest definition of the city is the city council boundary; although this definition does include a large portion of the Severn Estuary, west as far as the islands of Steep Holm and Flat Holm.^[3] A slightly less narrow definition is used by the Office for National Statistics; this includes built-up areas which adjoin Bristol but are not within the city council boundary, such as *Whitchurch* village, *Filton*, *Patchway*, *Bradley Stoke*, and excludes non-built-up areas within the city council boundary.^[4] The ONS has also defined an area which it calls the "Bristol Urban Area" which includes Kingswood, Mangotsfield, Stoke Gifford, Winterbourne, Frampton Cotterell, Almondsbury and Easton-in-Gordano.^[5] The term "Greater Bristol" (used for example by the Government Office of the South West ^[6]) is most usually used to refer to the area covered by the city and its three neighbouring local authorities.

Bristol



View from Cumberland Basin of the Clifton Suspension Bridge and the Avon Gorge



Coat of Arms of the City Council



Coordinates: 51°27′14″N 2°35′48″W﻿ / ﻿51.454°N 2.597°W﻿ / 51.454; -2.597

| | |
|--|--|
| Sovereign state | United Kingdom |
| Constituent country | England |
| Region | South West England |
| Ceremonial county | Bristol |
| Historic counties | County corporate (Gloucestershire and Somerset) <div>Avon</div> <div>Gloucestershire</div> |
| Royal Charter | 1155 |
| County status | 1373 |
| Government <div></div> | |
| - Type | Unitary authority, City |
| - Governing body | Bristol City Council |
| - Leadership | Leader & Cabinet |
| - Executive | Labour |
| - MPs | Roger Berry (L) <div>Kerry McCarthy (L)</div> <div>Doug Naysmith (L)(Co-op)</div> <div>Dawn Primarolo (L)</div> <div>Stephen Williams (LD)</div> |

PROJECT
HALO

Extracting Structured Data from Wikipedia

Title → **Bristol**

Description → **Bristol** (pronunciation (help·info); IPA: ˈbrɪstəl) is a city, unitary authority and ceremonial county in South West England, 105 miles (169 km) west of London, and 44 miles (71 km) east of Cardiff.

Domain specific Data → **Coordinates:** 51°28′N 2°35′W

Images → 
New from Clifton down to the Clifton Suspension Bridge and the Avon Gorge


Coat of Arms of the City Council

Infobox Properties →

| | |
|----------------------------|--|
| Sovereign state | United Kingdom |
| Constituent country | England |
| Region | South West England |
| Ceremonial county | Bristol |
| Historic counties | County corporate (Gloucestershire and Somerset) |
| Admin HQ | Bristol |
| Royal Charter | 1155 |
| County status | 1373 |
| Government | |
| - Type | Unitary authority, City |
| - Governing body | Bristol City Council |
| - Leadership | Leader & Cabinet |
| - Executive | Labour |
| - MPs | Roger Berry (L) Kerry McCarthy (L) Doug Naysmith (L)(Co-op) Dawn Primarolo (L) Stephen Williams (LD) |

Languages → **Anglo Saxon**
العربية
Български
Català
Česky
Cymraeg
Dansk
Deutsch
Eesti
Español
Esperanto
Euskara
Français
Ido
Bahasa Indonesia
Isleenska
Italiano
עברית
Latviešu
Lëtzebuergesch
Lietuvių
Magyar
Nederlands
日本語
Norsk (bokmål)

Further Down → **Web Links** → **Category**

Extracting Data from Wikipedia Tables

Table
Data

Porsche Cayenne - Wikipedia - Mozilla Firefox

http://wikivulcan.com/wp/index.php/Porsche_Cayenne

Porsche Cayenne - UltraPedia

Engines

| Model | Engine | Power (hp, torque)@rpm |
|--------------------------------------|--|---|
| Cayenne | 5,598 cc (5.598 L; 341.6 cu in) V6 | 290 PS (210 kW; 290 hp)@6200, 385 N·m (284 lb·ft)@3000 |
| Cayenne S | 4,806 cc (4.806 L; 293.3 cu in) V8 | 385 PS (283 kW; 380 hp)@6200, 500 N·m (370 lb·ft)@3500 |
| Cayenne S Transsyberia | 4,806 cc (4.806 L; 293.3 cu in) V8 | 405 PS (298 kW; 399 hp)@6500, 500 N·m (370 lb·ft)@3500 |
| Cayenne GTS | 4,806 cc (4.806 L; 293.3 cu in) V8 | 405 PS (298 kW; 399 hp)@6500, 500 N·m (370 lb·ft)@3500 |
| Cayenne GTS Porsche Design Edition 3 | 4,806 cc (4.806 L; 293.3 cu in) V8 | 405 PS (298 kW; 399 hp)@6500, 500 N·m (370 lb·ft)@3500 |
| Cayenne Turbo | 4,806 cc (4.806 L; 293.3 cu in) turbo V8 | 500 PS (370 kW; 490 hp)@6000, 700 N·m (520 lb·ft)@2250-4500 |
| Cayenne Turbo S | 4,806 cc (4.806 L; 293.3 cu in) twin turbo V8 | 550 PS (400 kW; 540 hp)@6000, 750 N·m (550 lb·ft)@2250-4500 |
| Cayenne Diesel | 2,967 cc (2.967 L; 181.1 cu in) turbo V6 | 240 PS (180 kW; 240 hp), 550 N·m (410 lb·ft)@2000 |
| Cayenne S Hybrid | 3.0L supercharged V6, 3-phase synchronous electric motor | Petrol: 333 PS (245 kW; 328 hp), 439 N·m (324 lb·ft)@2900-5300 Electric: 52 PS (38 kW; 51 hp), 300 N·m (220 lb·ft) |

6-speed automatic


Wheelbase: 2,855 mm (112.4 in)

Length: 4,798 mm (188.9 in)

Turbo: 4,795 mm (188.8 in)

Width: 1,928 mm (75.9 in)

Height: 1,700 mm (66.9 in)



Electronic Stabilizer decoupling demonstrating extra wheel articulation.

Second generation

The 2nd-generation Porsche Cayenne is expected to go on sale around April-May 2010 as a 2011 model, with an official debut at the 2010 *Geneva Motor Show*. In preparation for the upcoming unveiling, the Cayenne production facility in Leipzig, Germany, closed in December 2009 in order to commence the obligatory factory retooling for the new model, a process which is planned to take 2-3 months.

The first spy photos of the car were posted on the internet on the 5th of June 2008.^[9] Further spy photos, taken on June 2, 2009^[10] and between July 2009 and January 2010 reveal a shorter, smaller Cayenne with more muscular curves, a more slanted rear window and less upright windshield, a more sloping roofline, door-mounted mirrors, smaller windows at the rear of the vehicle, headlights inspired by the *Camerra GT*, taillights that extend onto the car's taillgate, LED *daytime running lights* and a vastly redesigned interior modeled after the *Panamera*. The new Cayenne is expected to be almost 250 kilograms lighter than the current model due to extensive use of aluminum and magnesium, making it more fuel efficient than the current lineup, as well as 5 centimeters shorter than the outgoing model. Due to its lower stance, however, the vehicle's off-road capabilities will be greatly reduced for a more performance-oriented layout and design. Diesel and hybrid variants will also be offered.

The Cayenne will again be the first of the three new SUVs from the VW group; the new *Volkswagen Touareg* will be 6–12 months behind, while the next-generation *Audi Q7* is due in 2013. Rumored standard features of the 2011 Porsche Cayenne will include air conditioning w/dual-zone climate controls, interior air filter, tilt/telescopic leather-wrapped steering wheel w/radio controls, cruise control, leather upholstery, 12-way power front seats, heated front seats, outside-temperature indicator, universal garage door opener, power liftgate, and power sunroof in the base model. The Cayenne S will add on tri-zone climate controls, heated steering wheel, and a compass. The Cayenne GTS will add on a rearview camera, remote engine start, keyless access and start, and memory system. Finally, the most upscale Cayenne Turbo and Turbo S will add on a navigation system w/voice recognition, premium sound system, 4-zone climate controls, heated rear seats, and 6-disc CD changer.^[11] The new Cayenne models will also offer Porsche's new Porsche Doppelkupplung (PDK) seven-speed *dual clutch transmission* instead of the currently-used six-speed Tiptronic S.

The Cayenne's engines are expected to receive a tuning upgrade, resulting in faster acceleration times with more horsepower and torque, as well as more powerful direct injection technology to improve efficiency. It is expected to source its V8 engines from the *Panamera*.

| Model | Engine | 0–60 mph | Price (USD) |
|-----------------|---------------------|-------------|-------------|
| Cayenne | 310 hp (231 kW) V6 | 7.0 seconds | \$ 47,000 |
| Cayenne S | 418 hp (312 kW) V8 | 6.0 seconds | \$ 60,000 |
| Cayenne GTS | 460 hp (343 kW) V8 | 5.2 seconds | \$ 80,000 |
| Cayenne Turbo | 533 hp (397 kW) V8 | 4.8 seconds | \$ 110,000 |
| Cayenne Turbo S | 580 hp (433 kW) V10 | 4.5 seconds | \$ 135,000 |

Porsche Cayenne 2nd generation

Production: 2010–

Engine(s): 310 hp (231 kW) V6
418 hp (312 kW) V8 (S)
460 hp (343 kW) V8 (GTS)
533 hp (397 kW) V8 (Turbo)
580 hp (430 kW) V10 (Turbo S)

Transmission(s): 6-speed manual
6-speed automatic

See also

3 List of hybrid vehicles

Data from Wikipedia: The DBpedia Project

■ The DBpedia Project (www.dbpedia.org)

- Publicly-funded EU Project to extract the knowledge collected by the Wikipedia community
 - Core is the Wikipedia Infobox data
 - ~700 infobox types, ~2800 property types
- Creates a database of structured information available on the web under an open-source license



■ DBpedia 3.5 dataset (Mar 10 Wikipedia)

- ~3.4M things, ~1B triples, 92 languages
- 312k persons, 413k places, 94k music albums, 49k films, 841k links to images, 5.1M links to relevant external web pages, 9.4M links into RDF datasets
- 3 Classification hierarchies

■ A large and broad knowledge base in the world



DBpedia Helps On Analytics

Skyscrapers in China
higher than 50 stories
built before year 2000

```
{ {#ask:
  [[Category:Skyscrapers]]
  [[Located in::China]]
  [[Floor count::>50]]
  [[Year built::<2000]]
  ... (output format)
} }
```



Jin Mao Tower

The Jin Mao Tower is an 88-story landmark supertall skyscraper in the Lujiazui area of the Pudong district of Shanghai, People's Republic of China. It contains offices and the Shanghai Grand Hyatt hotel. Until 2007 it was the tallest building in the PRC, the fifth tallest in the world by roof height and the seventh tallest by pinnacle height. Along with the Oriental Pearl Tower, it is a centerpiece of the Pudong skyline.



Hopewell Centre, Hong Kong

Hopewell Centre is a skyscraper in Hong Kong. It is located at 183 Queen's Road East, in Wan Chai on Hong Kong Island. It is the first circular skyscraper in Hong Kong. It is named after Hong Kong-listed property firm Hopewell Holdings Limited (合和實業有限公司), which constructed the building. Hopewell Holdings Limited is headquartered in the building and its CEO, Sir Gordon Wu Ying Sheung, has his office on the top floor of the building.



Shun Hing Square

Shun Hing Square is a 384m (1,260 ft) tall skyscraper in Shenzhen, China. It is the tallest steel building in China, the 5th tallest building in China, and the 9th tallest in the world. It was the tallest building in China from its completion in 1996 until CITIC Plaza in Guangzhou was completed the next year. The building was built at the fast pace of four floors in nine days.



Highcliff

Highcliff is a 252.4-metre (828-foot) tall skyscraper located on a south slope of Happy Valley on the Hong Kong Island in Hong Kong. The 75 storey (70 floors of which are livable space) building's construction began in 2000 and was completed in 2003 under a design by DLN Architects & Engineers. It was the Silver Winner of the 2003 Emporis Skyscraper Award, coming in second to 30 St Mary Ave in London.



The Harbourside

The Harbourside is a 255 m (836.6 ft) tall residential skyscraper located at 1 Austin Road West, in Union Square complex on Kowloon peninsula. The building is erected on the West Kowloon Reclamation west of Kwun Chung. Construction of the 74 storey building began in 2000 and was completed in 2003 under the design by P & T Architects & Engineers. The building is, in fact, three towers joined at the base, middle and top however from a distance it appears as one wall of a building.

Project Halo Enhancements to DBpedia

■ Enhanced Wikipedia parsing support

- Tables into semantics
- Multiple infoboxes per page; cleaner infobox processing
- Provenance (Wikipedia source line number) for every assertion

■ Live update stream and on-demand extraction API

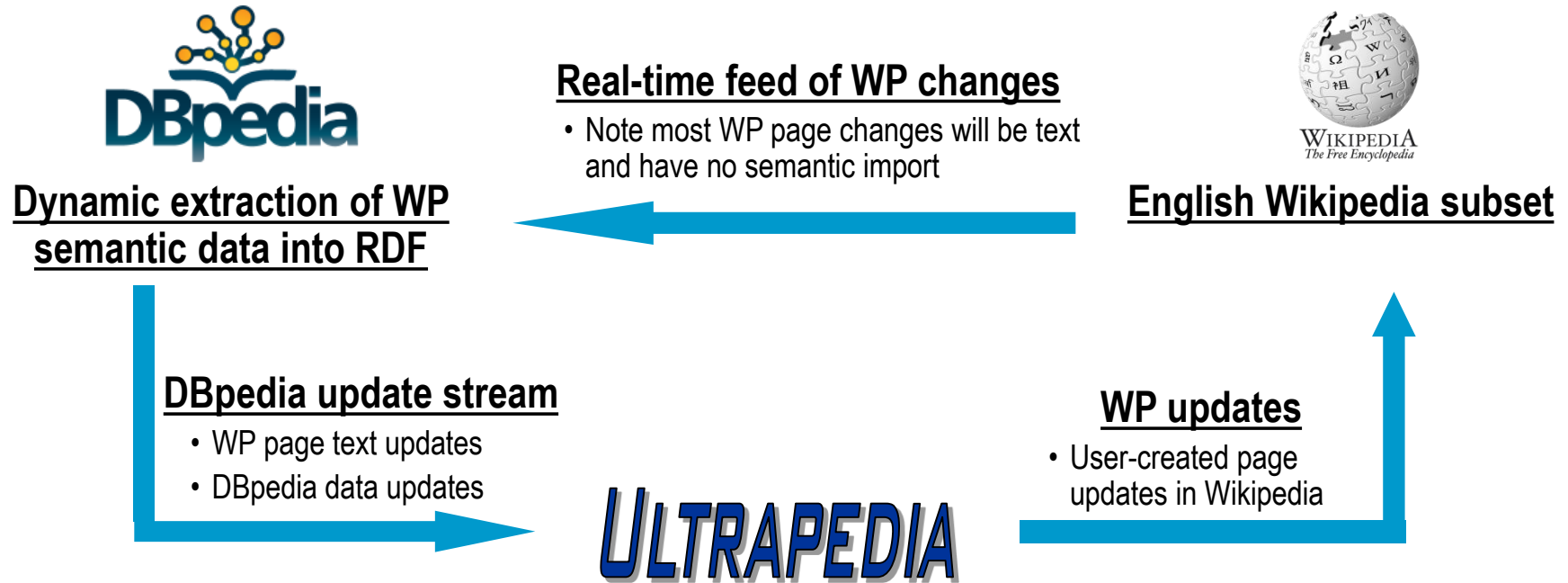
■ Data structure definition and template mapping

- Define an SWM ontology for template/table data
- Data quality improvements
 - Validate input data against expected ranges, units, dimensions (“born” as a date, a city, or both?)
 - Ability to convert units (meters to kilometers to miles etc.)
- Mappings maintained on <http://mapping.dbpedia.org>
 - Class and property definitions
 - Mappings from Wikipedia templates/tables to ontology classes
 - Mappings from Wikipedia template properties and table columns to ontology properties

■ Numerous bug fixes and tweaks



Ultrapedia Prototype Data Flow



Enhanced Ultrapedia Usability

- Familiar WP page text and layout
- Exhibit-based visualizations
- Dynamic tables/categories
- Faceted navigation
- Queries (both standing and ad-hoc)
- Wikitag-based MS Office augmentation

Wikipedia-based Corrections

- UP shows the user where to correct data in WP so that DBpedia will extract the correction
 - Ultrapedia exposes the data source in terms of where the data was extracted from WP
- WP changes and corrections get quickly propagated to UP



Ultrapedia Demo

■ Domain is German cars

- Cars, Companies, Engines, Transmissions, People, etc.
- ~2500 pages, ~40000 triples

■ An SMW-based encyclopedia

- Similar look and feel to Wikipedia
- With dynamic tables
- Flexible and powerful queries
- Many visualization methods
- Trustworthy data source
- Edit, discuss and rate *data*

■ Things to take away

- A better Wikipedia
- Some pages are Ultrapedia only
- External data integration via web services (EBay)
- Real-time data updates

■ Rapid to build

- Software is quite stable
- Most time was spent on data cleaning and new visualizations

Next Steps for Ultrapedia

- **Even better data extraction system**
- **Needs a natural, easy-to-use query system**
- **Wikipedia for neuroscience knowledge is next**
 - Size is about 10x German cars
- **We will also integrate with other Project Halo components**
- **What would it take to scale up to Wikipedia?**
 - ~3M English articles, ~1M German articles, long tail of 264 languages
 - Key scaling factors are:
 - Table/infobox mappings: We have opened these to crowdsourcing
 - Triplestore query times: currently ~50ms for ~40K triples, we have room to grow by a factor of 100-1000
 - Fixing data errors and table parse errors in Wikipedia
 - This is a manual process that is done per article



SMW Contributors

Paul Allen
Free University of Berlin
John Chu
Mark Greaves
Benjamin Grosof
Dave Gunning
Rob Kaplan
Yaron Koren
University of Karlsruhe
Dave Moore
Ontoprise GmbH
TeamMersion LLC
Jesse Wang



Thank You

